

# Tracking Switched Dynamic Network Topologies from Information Cascades<sup>†</sup>

Brian Baingana, Student Member, IEEE and Georgios B. Giannakis, Fellow, IEEE\*

**Abstract**—Contagions such as the spread of popular news stories, or infectious diseases, propagate in cascades over dynamic networks with unobservable topologies. However, “social signals” such as product purchase time, or blog entry timestamps are measurable, and implicitly depend on the underlying topology, making it possible to track it over time. Interestingly, network topologies often “jump” between discrete states that may account for sudden changes in the observed signals. The present paper advocates a switched dynamic structural equation model to capture the topology-dependent cascade evolution, as well as the discrete states driving the underlying topologies. Conditions under which the proposed switched model is identifiable are established. Leveraging the edge sparsity inherent to social networks, a recursive  $\ell_1$ -norm regularized least-squares estimator is put forth to jointly track the states and network topologies. An efficient first-order proximal-gradient algorithm is developed to solve the resulting optimization problem. Numerical experiments on both synthetic data and real cascades measured over the span of one year are conducted, and test results corroborate the efficacy of the advocated approach.

**Index Terms**—Social networks, structural equation model, network cascade, topology inference, switched linear systems.

## I. INTRODUCTION

Information often spreads in cascades by following implicit links between nodes in real-world networks whose topologies may be unknown. For example the spread of viral news among blogs, or Internet memes over microblogging tools like *Twitter*, are facilitated by the inherent connectivity of the world-wide web. A celebrity may post an interesting *tweet*, which is then read and *retweeted* by her followers. An information cascade may then emerge if the followers of her followers share the tweet, and so on. The dynamics of propagation of such information over implicit networks are remarkably similar to those that govern the rapid spread of infectious diseases, leading to the so-termed *contagions* [1], [7], [21]. Similar cascading processes have also been observed in the context of adoption of emerging fashion trends within distinct age groups, the successive firing of thousands of neurons in the

brain in response to stimuli, and plummeting stock prices in global financial markets in response to a natural disaster.

Cascades are generally observed by simply recording the time when a specific website first mentioned a cascading news item, or when an infected person first showed symptoms of a disease. On the other hand, the underlying network topologies may be unknown and dynamic, with the link structure varying over time. Unveiling such dynamic network topologies is crucial for several reasons. Viral web advertising can be more effective if a small set of influential early adopters are identified through the link structure, while knowledge of the structure of hidden needle-sharing networks among communities of injecting drug users can aid formulation of policies for curbing contagious diseases. Other examples include assessment of the reliability of heavily interconnected systems like power grids, or risk exposure among investment banks in a highly inter-dependent global economy. In general, knowledge of topologies that facilitate diffusion of network processes leads to useful insights about the behavior of complex systems.

Network contagions arise due to *causal* interactions between nodes e.g., blogs, or disease-susceptible individuals. *Structural equation models* (SEMs) effectively capture such causal relationships, that are seldom revealed by symmetric correlations; see e.g., [10], [18]. Widely applied in psychometrics [15], and sociometrics [8], SEMs have also been adopted for gene network inference [5], [12], and brain connectivity studies [13]. Recently, dynamic SEMs have been advocated for tracking slowly-varying sparse social network topologies from cascade data [1].

Network topologies may sometimes “jump” between a finite number of discrete states, as manifested by sudden changes in cascade behavior. For example, an e-mail network may switch topologies from predominantly work-based connections during the week, to friend-based connections over the weekend. Connection dynamics between bloggers may switch suddenly at the peak of sports events (e.g., “Superbowl”), or presidential elections. In such settings, contemporary approaches assuming that network dynamics arise as a result of slow topology variations may yield unpredictable results. The present paper capitalizes on this prior knowledge, and puts forth a novel *switched* dynamic SEM to account for propagation of information cascades in such scenarios. The novel approach builds upon the dynamic SEM advocated in [1], where it is tacitly assumed that node infection times depend on both the switching topologies and exogenous influences e.g., external information sources, or prior predisposition to certain cascades.

The present paper draws connections to identification of hybrid systems, whose behavior is driven by interaction between continuous and discrete dynamics; see e.g., [16] and references

<sup>†</sup> Work in this paper was supported by NSF grants 1343248, 1442686, 1514056, NIH Grant No.1R01GM104975-01, and AFOSR-MURI Grant No. FA9550-10-1-0567. This research has been co-financed by the European Union (European Social Fund - ESF), and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the ESF. Parts of the paper will appear in the *Proc. of the IEEE Global Conference on Signal and Information Processing*, Orlando, Florida, December 14-16, 2015

\* The authors are with the Dept. of ECE and the Digital Technology Center, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455. Tel/fax: (612)626-7781/625-4583; Emails: {baing011, georgios}@umn.edu

therein. Switched linear models have emerged as a useful framework to capture piecewise linear input-output relations in control systems [2], [22], [23]. The merits of these well-grounded approaches are broadened here to temporal network inference from state-driven cascade dynamics. Although the evolution of the unknown network state sequence may be controlled by structured hidden dynamics (e.g., *hidden Markov models*), this work advocates a more general framework in which such prior knowledge is not assumed.

Network inference from temporal traces of infection events has recently emerged as an active research area. A number of approaches put forth probabilistic models for information diffusion, and leverage maximum likelihood estimation (MLE) to infer both static and dynamic edge weights as pairwise transmission weights between nodes [14], [19], [20]. Sparse SEMs are leveraged to capture exogenous inputs in inference of dynamic social networks in [1], and static causal links in gene regulatory networks [5]. Network Granger causality with group sparsity is advocated for inference of causal networks with inherent grouping structure in [3], while causal influences are inferred by modeling historical network events as multidimensional Hawkes processes in [25].

Within the context of prior works on dynamic network inference, the contributions of the present paper are three-fold. First, a novel switched dynamic SEM that captures sudden topology changes within a finite state-space is put forth (Section II). Second, identifiability results for the proposed switched model are established under reasonable assumptions, building upon prior results for static causal networks in [4] (Section III). Finally, an efficient sparsity-promoting proximal-gradient (PG) algorithm is developed (Sections IV and V) to jointly track the evolving state sequence, and the unknown network topologies. Numerical tests on both synthetic and real cascades in Section VI corroborate the efficacy of the novel approach. Interestingly, experiments on real-world web cascade data exemplify that media influence is dominated by major news outlets (e.g., *cnn.com* and *bbc.com*), as well as well-known web-based news aggregators (e.g., *news.yahoo.com* and *news.google.com*).

*Notation.* Bold uppercase (lowercase) letters will denote matrices (column vectors), while operators  $(\cdot)^\top$ ,  $\lambda_{\max}(\cdot)$ , and  $\text{diag}(\cdot)$  will stand for matrix transposition, maximum eigenvalue, and diagonal matrix, respectively. The identity matrix will be represented by  $\mathbf{I}$ , while  $\mathbf{0}$  will denote the matrix of all zeros, and their dimensions will be clear in context. Finally, the  $\ell_p$  and Frobenius norms will be denoted by  $\|\cdot\|_p$ , and  $\|\cdot\|_F$ , respectively.

## II. MODEL AND PROBLEM STATEMENT

Consider a dynamic network with  $N$  nodes observed over time intervals  $t = 1, \dots, T$ , captured by a graph whose topology may arbitrarily switch between  $S$  discrete states. Suppose the network topology during interval  $t$  is described by an unknown, weighted adjacency matrix  $\mathbf{A}^{\sigma(t)} \in \mathbb{R}^{N \times N}$ , with  $\sigma(t) \in \mathcal{S} := \{1, \dots, S\}$ . Entry  $(i, j)$  of  $\mathbf{A}^{\sigma(t)}$  (henceforth denoted by  $a_{ij}^{\sigma(t)}$ ) is nonzero only if a directed edge connects nodes  $i$  and  $j$  (pointing from  $j$  to  $i$ ) during interval  $t$ . In general  $a_{ij}^{\sigma(t)} \neq a_{ji}^{\sigma(t)}$ , i.e., matrix  $\mathbf{A}^{\sigma(t)}$  is generally non-symmetric,

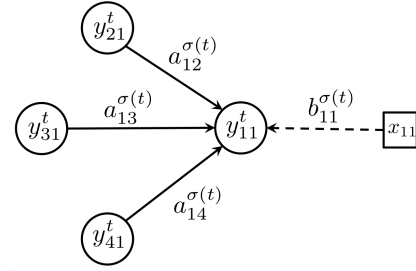


Fig. 2: The cascade infection time for node 1 during interval  $t$  is a linear combination of cascade infection times of its single-hop neighbors, and the exogenous influence  $x_{11}$ . Unknown edge weights are captured through the coefficients  $a_{12}^{\sigma(t)}$ ,  $a_{13}^{\sigma(t)}$  and  $a_{14}^{\sigma(t)}$ .

which is suitable to model directed networks. The model tacitly assumes that the network topology remains fixed during any given time interval  $t$ , but can change across time intervals.

Suppose  $C$  contagions propagating over the network are observed, and the time of first infection per node  $i$  by contagion  $c$  is denoted by  $y_{ic}^t \in \mathbb{R}$ . In online media,  $y_{ic}^t$  can be obtained by recording the timestamp when blog  $i$  first mentioned news item  $c$ . If a node remains uninfected during  $t$ ,  $y_{ic}^t$  is set to an arbitrarily large number. Assume that the susceptibility  $x_{ic}$  of node  $i$  to external (non-topological) infection by contagion  $c$  is known and time invariant. In the web context,  $x_{ic}$  can be set to the search engine rank of website  $i$  with respect to (w.r.t.) keywords associated with  $c$ .

The infection time of node  $i$  during interval  $t$  is modeled according to the following *switched* dynamic structural equation model (SEM)

$$y_{ic}^t = \sum_{j \neq i} a_{ij}^{\sigma(t)} y_{jc}^t + b_{ii}^{\sigma(t)} x_{ic} + e_{ic}^t \quad (1)$$

where  $b_{ii}^{\sigma(t)}$  captures the state-dependent level of influence of external sources, and  $e_{ic}^t$  accounts for measurement errors and unmodeled dynamics. It follows from (1) that if  $a_{ij}^{\sigma(t)} \neq 0$ , then  $y_{ic}^t$  is affected by the values of  $\{y_{jc}^t\}_{j \neq i}$  (see Figure 2). With diagonal  $\mathbf{B}^{\sigma(t)} := \text{diag}(b_{11}^{\sigma(t)}, \dots, b_{NN}^{\sigma(t)})$ , collecting observations for the entire network and all  $C$  contagions yields the dynamic matrix SEM

$$\mathbf{Y}_t = \mathbf{A}^{\sigma(t)} \mathbf{Y}_t + \mathbf{B}^{\sigma(t)} \mathbf{X} + \mathbf{E}_t \quad (2)$$

where  $\mathbf{Y}_t := [y_{ic}^t]$ ,  $\mathbf{X} := [x_{ic}]$ , and  $\mathbf{E}_t := [e_{ic}^t]$  are all  $N \times C$  matrices. A single network topology  $\mathbf{A}^{\sigma(t)}$  is adopted for all contagions, which is suitable e.g., when information cascades are formed around a common meme or trending (news) topic over the Internet. Matrix  $\mathbf{A}^{\sigma(t)}$  has an all-zero diagonal to ensure that the underlying network is free of self loops, i.e.,  $a_{ii}^{\sigma(t)} = 0 \ \forall i, \forall t$ .

**Problem statement.** Given the sequence  $\{\mathbf{Y}_t\}_{t=1}^T$  and  $\mathbf{X}$  adhering to (2), the goal is to identify the unknown state matrices  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$ , and the switching sequence  $\{\sigma(t) \in \mathcal{S}\}_{t=1}^T$ .

## III. MODEL IDENTIFIABILITY

This section explores the conditions under which one can uniquely recover the unknown state matrices  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$ .

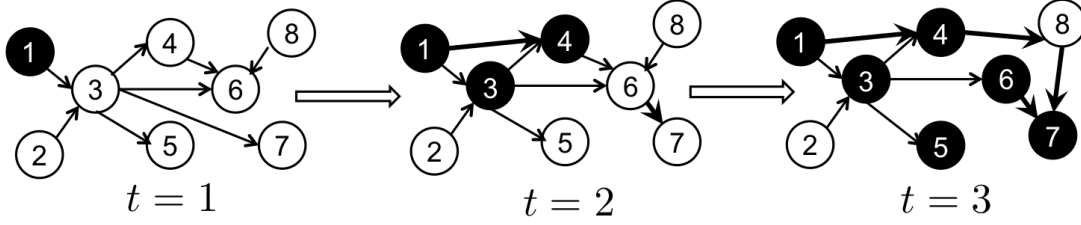


Fig. 1: A single cascade propagating over a time-varying network, with “infected” nodes depicted in black. Some edges disappear over consecutive time intervals as new ones emerge (thick arrows). Time-varying changes to the network topology affect the progression of the cascade over the network.

First, (2) can be written as

$$\mathbf{Y}_t = \left( \sum_{s=1}^S \chi_{ts} \mathbf{A}^s \right) \mathbf{Y}_t + \left( \sum_{s=1}^S \chi_{ts} \mathbf{B}^s \right) \mathbf{X} + \mathbf{E}_t \quad (3)$$

where the binary indicator variable  $\chi_{ts} = 1$  if  $\sigma(t) = s$ , otherwise  $\chi_{ts} = 0$ . Note that (3) is generally underdetermined when  $S$  is unknown, and it may not be possible to uniquely identify  $\{\mathbf{A}^s, \mathbf{B}^s, \{\chi_{ts}\}_{t=1}^T\}_{s=1}^S$ . Even in the worst-case scenario where  $S = T$ , complete identification of all states is impossible if there exists  $t$  and  $t'$  ( $t \neq t'$ ) such that  $\sigma(t) = \sigma(t')$ . In order to establish model identifiability, it will be assumed that the following hold:

**as1.** The dynamic SEM in (2) is noise-free ( $\mathbf{E}_t = \mathbf{0}$ ); that is,

$$\mathbf{Y}_t = \mathbf{A}^{\sigma(t)} \mathbf{Y}_t + \mathbf{B}^{\sigma(t)} \mathbf{X}. \quad (4)$$

**as2.** All cascades  $\{\mathbf{Y}_t\}_{t=1}^T$  are generated by some pair  $\{\mathbf{A}^s, \mathbf{B}^s\}$ , where  $s \in \{1, \dots, S\}$ , and  $S$  is known. This is a *realizability assumption*, which guarantees the existence of a submodel responsible for the observed cascades.

**as3.** No two states can be active during a given time interval, i.e.,  $\|\mathbf{Y}_t - \mathbf{A}^s \mathbf{Y}_t - \mathbf{B}^s \mathbf{X}\|_F = \|\mathbf{Y}_t - \mathbf{A}^{s'} \mathbf{Y}_t - \mathbf{B}^{s'} \mathbf{X}\|_F = 0$  implies that  $s = s'$ .

Under (as1)-(as3), the following proposition holds.

**Proposition 1:** Suppose data matrices  $\mathbf{Y}_t$  and  $\mathbf{X}$  adhere to (4) with  $a_{ii}^{\sigma(t)} = 0$ ,  $b_{ii}^{\sigma(t)} \neq 0$ ,  $\forall i, t$ , and  $b_{ij}^{\sigma(t)} = 0 \forall i \neq j, t$ . If  $N \leq C$  and  $\mathbf{X}$  has full row rank, then  $\mathbf{A}^{\sigma(t)}$  and  $\mathbf{B}^{\sigma(t)}$  are uniquely expressible in terms of  $\mathbf{X}$  and  $\mathbf{Y}_t$  as  $\mathbf{B}^{\sigma(t)} = (\text{Diag}[(\mathbf{Y}_t \mathbf{X}^\dagger)])^{-1}$ , and  $\mathbf{A}^{\sigma(t)} = \mathbf{I} - \mathbf{B}^{\sigma(t)} (\mathbf{Y}_t \mathbf{X}^\dagger)^{-1}$ .

If  $p_s := \Pr(\chi_{ts} = 1) > 0 \forall s$  denotes the activation probability of state  $s$ , with  $\sum_{s=1}^S p_s = 1$ , and (as1)-(as3) hold, then  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  can be uniquely identified with probability one as  $T \rightarrow \infty$ .

Proposition 1 establishes a two-step identifiability result for the dynamic SEM model in (4). First, it establishes that if the exogenous data matrix  $\mathbf{X}$  is sufficiently rich (i.e.,  $\mathbf{X}$  has full row rank), then the per-interval network topology captured by  $\mathbf{B}^{\sigma(t)}$  can be uniquely identified per  $t$ . Second, if the state activation probabilities are all strictly positive, then the  $S$  switching network topologies will all be recovered as  $T \rightarrow \infty$ . In fact, if cascade data are acquired in an infinitely streaming fashion ( $T = \infty$ ), then one is guaranteed to uniquely recover all states.

**Proof of Proposition 1:** Equation (4) can be written as  $(\mathbf{I} - \mathbf{A}^{\sigma(t)}) \mathbf{Y}_t = \mathbf{B}^{\sigma(t)} \mathbf{X}$ , which implies that

$\text{rank}((\mathbf{I} - \mathbf{A}^{\sigma(t)}) \mathbf{Y}_t) = \text{rank}(\mathbf{B}^{\sigma(t)} \mathbf{X})$ . With  $\mathbf{X}$  and  $\mathbf{B}^{\sigma(t)}$  both having full row rank (recall that  $b_{ii}^{\sigma(t)} > 0$ ), it follows that  $(\mathbf{I} - \mathbf{A}^{\sigma(t)}) \mathbf{Y}_t$  has full row rank, and  $\mathbf{I} - \mathbf{A}^{\sigma(t)}$  is invertible. Consequently, the linear system of equations  $\mathbf{Y}_t = \Phi \mathbf{X}$  is solved by  $\Phi^* = (\mathbf{I} - \mathbf{A}^{\sigma(t)})^{-1} \mathbf{B}^{\sigma(t)}$ .

On the other hand, note that  $\mathbf{Y}_t = \Phi \mathbf{X}$  admits the solution

$$\Phi^* = \mathbf{Y}_t \mathbf{X}^\dagger \quad (5)$$

which is unique since  $\mathbf{X}$  is full row rank. Since  $a_{ii}^{\sigma(t)} = 0$  and  $(\mathbf{B}^{\sigma(t)})^{-1}$  is a diagonal matrix, the diagonal entries of  $(\Phi^*)^{-1} = (\mathbf{B}^{\sigma(t)})^{-1} (\mathbf{I} - \mathbf{A}^{\sigma(t)})$  coincide with those of  $(\mathbf{B}^{\sigma(t)})^{-1}$ ; hence,  $\mathbf{B}^{\sigma(t)} = (\text{Diag}[(\Phi^*)^{-1}])^{-1}$ , which leads to

$$\mathbf{B}^{\sigma(t)} = (\text{Diag}[(\mathbf{Y}_t \mathbf{X}^\dagger)^{-1}])^{-1} \quad (6)$$

upon substitution of (5). Furthermore, note that  $\mathbf{B}^{\sigma(t)} (\mathbf{Y}_t \mathbf{X}^\dagger)^{-1} = \mathbf{I} - \mathbf{A}^{\sigma(t)}$ , and thus

$$\mathbf{A}^{\sigma(t)} = \mathbf{I} - [\text{Diag}[(\mathbf{Y}_t \mathbf{X}^\dagger)^{-1}]]^{-1} (\mathbf{Y}_t \mathbf{X}^\dagger)^{-1} \quad (7)$$

which concludes the first part of the proof.

Since  $\mathbf{A}^{\sigma(t)} = \sum_{s=1}^S \chi_{ts} \mathbf{A}^s$  and  $\mathbf{B}^{\sigma(t)} = \sum_{s=1}^S \chi_{ts} \mathbf{B}^s$ , the results from the unique solutions (6) and (7) coincide with a specific pair in the set  $\{(\mathbf{A}^s, \mathbf{B}^s)\}_{s=1}^S$  per  $t$ . Intuitively, complete recovery of all state matrices is tantamount to identification of  $S$  unique pairs  $(\mathbf{A}^s, \mathbf{B}^s)$  as more data are sequentially acquired. In order to guarantee identifiability of all states in the long run, it suffices to prove that the probability of activation of any state at least once tends to 1 as  $T \rightarrow \infty$ . Since  $\chi_{ts}$  are Bernoulli random variables with  $p_s = \Pr(\chi_{ts} = 1)$  per  $t$ , the number of times that state  $s$  is activated over  $T$  time intervals follows a binomial distribution. Letting  $T_s := \sum_{t=1}^T \chi_{ts}$  denote the total number of activations of state  $s$  over  $t = 1, \dots, T$ , then

$$\begin{aligned} \Pr(T_s \geq 1) &= 1 - \Pr(T_s = 0) \\ &= 1 - (1 - p_s)^T. \end{aligned} \quad (8)$$

Since  $\lim_{T \rightarrow \infty} \Pr(T_s \geq 1) = 1$  only if  $p_s > 0$ , it follows that all states can be uniquely identified with probability 1 as  $T \rightarrow \infty$  when  $p_s > 0 \forall s$ , which completes the second part of the proof.



**Algorithm 1** Topology identification by clustering

---

**Require:**  $\{\mathbf{Y}_t\}_{t=1}^T, \mathbf{X}, S, T_{\text{cluster}} < T$

- 1: **for**  $t = 1, \dots, T_{\text{cluster}}$  **do**
- 2:    $\mathbf{b}^{\sigma(t)} = \text{diag}[(\text{Diag}[(\mathbf{Y}_t \mathbf{X}^\top)^{-1}])^{-1}]$
- 3:    $\mathbf{A}^{\sigma(t)} = \mathbf{I} - [\text{Diag}[(\mathbf{Y}_t \mathbf{X}^\top)^{-1}]]^{-1} (\mathbf{Y}_t \mathbf{X}^\top)^{-1}$
- 4:    $\boldsymbol{\theta}^t := \text{vec}([\mathbf{A}^{\sigma(t)} \mathbf{b}^{\sigma(t)}])$
- 5: **end for**
- 6:  $\{\boldsymbol{\theta}^s, \{\chi_{ts}\}_{t=1}^{T_{\text{cluster}}}\}_{s=1}^S = \text{cluster}(\{\boldsymbol{\theta}^t\}_{t=1}^{T_{\text{cluster}}}, S)$
- 7: Set  $\hat{\sigma}(t) = s$  if  $\chi_{ts} = s$  for  $t = 1, \dots, T_{\text{cluster}}$
- 8: Extract  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  from  $\{\boldsymbol{\theta}^s\}_{s=1}^S$
- 9: **for**  $t > T_{\text{cluster}}$  **do**
- 10:   Compute  $\boldsymbol{\theta}^t$  using lines (2) – (4)
- 11:    $\hat{\sigma}(t) = \arg \min_s \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^s\|_2$
- 12: **end for**
- 13: **return**  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S, \{\hat{\sigma}(t)\}_{t=1}^T$

---

**A. Topology tracking by clustering when  $\mathbf{E}_t \neq \mathbf{0}$** 

In general, even when  $\mathbf{X}$  is full row rank,  $\mathbf{E}_t \neq \mathbf{0}$  in order to compensate for measurement errors and unmodeled dynamics. Under noisy conditions, it will turn out that  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  can be interpreted as cluster centroids (cf. (3)), and one can leverage traditional clustering approaches (e.g., k-means), coupled with the closed-form solutions in (6) and (7) to identify the unknown state matrices. Indeed, with  $\mathbf{b}^{\sigma(t)} := \text{diag}(\mathbf{B}^{\sigma(t)})$ , and  $\boldsymbol{\theta}^t := \text{vec}([\mathbf{A}^{\sigma(t)} \mathbf{b}^{\sigma(t)}])$ , it follows readily that

$$\boldsymbol{\theta}^t = \sum_{s=1}^S \chi_{ts} \boldsymbol{\theta}^s \quad t = 1, \dots, T \quad (9)$$

where  $\boldsymbol{\theta}^s := \text{vec}([\mathbf{A}^s \mathbf{b}^s])$ , and  $\mathbf{b}^s := \text{diag}(\mathbf{B}^s)$ .

Clearly,  $\{\boldsymbol{\theta}^s\}_{s=1}^S$  in (9) can be viewed as  $S$  cluster centers, and  $\chi_{ts}$  as unknown binary cluster-assignment variables. Consequently, the sequence  $\{\boldsymbol{\theta}^t\}_{t=1}^{T_{\text{cluster}}}$  can be directly computed per  $t$  via (6) and (7), where  $T_{\text{cluster}} < T$  denotes the number of training samples. Identification of  $\{\chi_{ts}\}$  and  $\{\boldsymbol{\theta}^s\}$  can then be accomplished by batch clustering  $\{\boldsymbol{\theta}^t\}$  into  $S$  clusters. The subsequent operational phase ( $t > T_{\text{cluster}}$ ) then boils down to computing  $\boldsymbol{\theta}^t$ , followed by finding the centroid  $\boldsymbol{\theta}^s$  to which it is closest in Euclidean distance, that is

$$\hat{s} = \arg \min_s \|\boldsymbol{\theta}^t - \boldsymbol{\theta}^s\|_2 \quad (10)$$

and  $\hat{\chi}_{ts} = 1$  if  $s = \hat{s}$ , otherwise  $\hat{\chi}_{ts} = 0$ . Algorithm 1 summarizes this cluster-based state identification scheme, with a clustering phase ( $t \leq T_{\text{cluster}}$ ) and an operational phase ( $t > T_{\text{cluster}}$ ). The sub-procedure  $\text{cluster}(\{\boldsymbol{\theta}^t\}_{t=1}^{T_{\text{cluster}}}, S)$  calls an off-the-shelf clustering algorithm with the training set  $\{\boldsymbol{\theta}^t\}_{t=1}^{T_{\text{cluster}}}$  and the number of clusters  $S$  as inputs.

**Remark 1:** Algorithm 1 identifies the unknown state-dependent matrices by resorting to “hard clustering,” which entails deterministic assignment of  $\boldsymbol{\theta}^t$  to one of the  $S$  centroids. In principle, the algorithm can be readily modified to adopt “soft clustering” approaches, with probabilistic assignments to the cluster centroids.

**IV. EXPLOITING EDGE SPARSITY**

The fundamental premise established by Proposition 1 is that  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  are uniquely identifiable provided  $\mathbf{X}$  is sufficiently rich. However, requiring that  $\mathbf{X}$  has full row rank is a restrictive condition, tantamount to requiring that at least as many cascades are observed as the number of nodes ( $C \geq N$ ). This is especially prohibitive in large-scale networks such as the world-wide web, with billions of nodes. It is therefore of interest to measure as few cascades as possible while ensuring that  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  are uniquely recovered. To this end, one is motivated to leverage prior knowledge about the unknowns in order to markedly reduce the amount of data required to guarantee model identifiability. For example, each node in the network is connected only to a small number of nodes out of the  $N - 1$  possible connections. As a result, most practical networks exhibit edge sparsity, a property that can be exploited to ensure that the rank condition on  $\mathbf{X}$  can be relaxed, as shown in the sequel. In addition to (as1)-(as3), consider the following.

**as4.** Each row  $\mathbf{a}_n^s$  of matrix  $\mathbf{A}^s$  has at most  $K$  nonzero entries; i.e.,  $\|\mathbf{a}_n^s\|_0 \leq K \quad \forall n, s$ .

**as5.** The nonzero entries of  $\mathbf{A}^s$  for all  $s = 1, \dots, S$  are drawn from a continuous distribution.

**as6.** The Kruskal rank of  $\mathbf{X}^\top$  satisfies  $\text{kr}(\mathbf{X}^\top) \geq 2K + 1$ , where  $\text{kr}(\mathbf{Z})$  is defined as the maximum number  $k$  such that any combination of  $k$  columns of  $\mathbf{Z}$  constitute a full column rank submatrix.

**Proposition 2:** Suppose data matrices  $\mathbf{Y}_t$  and  $\mathbf{X}$  adhere to (4) with  $a_{ii}^{\sigma(t)} = 0$ ,  $b_{ii}^{\sigma(t)} \neq 0$ ,  $\forall i, t$ , and  $b_{ij}^{\sigma(t)} = 0 \quad \forall i \neq j, t$ . If (as1)-(as6) hold, and  $p_s > 0 \quad \forall s$ , then  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  can be uniquely identified with probability one as  $T \rightarrow \infty$ .

The proof of Proposition 2 is rather involved, and is deferred to Appendix A. Unlike the matrix rank which only requires existence of a subset of linearly independent columns, the Kruskal rank requires that every possible combination of a given number of columns be linearly independent. Moreover, computing  $\text{kr}(\mathbf{X}^\top)$  is markedly more challenging, as it entails a combinatorial search over the rows of  $\mathbf{X}$ . Admittedly, requiring  $\text{kr}(\mathbf{X}^\top) \geq 2K + 1$  is more restrictive than  $\text{rank}(\mathbf{X}^\top) \geq 2K + 1$ . Nevertheless, in settings where  $2K + 1 \ll N$ , (as6) may be satisfied even if  $\text{rank}(\mathbf{X}) < N$ . In such cases, Proposition 2 asserts that one can uniquely identify  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  even when  $C \leq N$ .

**A. Sparsity-promoting estimator**

The remainder of the present paper leverages inherent edge sparsity to develop an efficient algorithm to track switching network topologies from noisy cascade traces. Assuming  $S$  is known a priori, one is motivated to solve the following regularized LS batch estimator

$$\begin{aligned} \text{(P0)} \quad & \arg \min_{\substack{\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S \\ \{\chi_{ts}\}_{s=1}^S \mathbf{Y}_t}} \frac{1}{2} \sum_{t=1}^T \sum_{s=1}^S \chi_{ts} \|\mathbf{Y}_t - \mathbf{A}^s \mathbf{Y}_t - \mathbf{B}^s \mathbf{X}\|_F^2 \\ & + \sum_{s=1}^S \lambda_s \|\mathbf{A}^s\|_1 \end{aligned}$$

$$\begin{aligned} \text{s. to } \quad & \sum_{s=1}^S \chi_{ts} = 1 \quad \forall t, \quad \chi_{ts} \in \{0, 1\} \quad \forall s, t \\ & a_{ii}^s = 0, \quad b_{ij}^s = 0, \quad \forall s, i \neq j \end{aligned} \quad (11)$$

where the constraint  $\sum_{s=1}^S \chi_{ts} = 1$  enforces a realizability condition, ensuring that only one state can account for the system behavior at any time. With  $\|\mathbf{A}^s\|_1 := \sum_{ij} |a_{ij}^s|$ , the regularization term promotes edge sparsity that is inherent to most real networks. The sparsity level of  $\hat{\mathbf{A}}^s$  is controlled by  $\lambda_s > 0$ . Absence of a self-loop at node  $i$  is enforced by the constraint  $a_{ii}^s = 0$ , while having  $b_{ij}^s = 0, \forall i \neq j$ , ensures that  $\hat{\mathbf{B}}^s$  is diagonal as in (2).

Note that (P0) is an NP-hard mixed integer program that is unsuitable for large-scale and potentially real-time operation. Moreover, entries of per-state matrices  $\{\mathbf{A}^s, \mathbf{B}^s\}$  may evolve slowly over reasonably long observation periods, motivating algorithms that not only track the switching sequence, but also the slow drifts occurring in per-state network topologies.

Suppose data are sequentially acquired, rendering batch estimators impractical. If it is assumed that instantaneous and past estimates of  $\{\{\hat{\chi}_{\tau s}\}_{s=1}^S\}_{\tau=1}^t$  are known during interval  $t$ , (P0) decouples over  $S$ , and is tantamount to solving the following subproblem per  $t$  and  $s$

$$\begin{aligned} \text{(P1)} \quad & \arg \min_{\mathbf{A}^s, \mathbf{B}^s} \quad \frac{1}{2} \sum_{\tau=1}^t \hat{\chi}_{\tau s} \|\mathbf{Y}_\tau - \mathbf{A}^s \mathbf{Y}_\tau - \mathbf{B}^s \mathbf{X}\|_F^2 \\ & + \lambda_s \|\mathbf{A}^s\|_1 \\ \text{s. to} \quad & a_{ii}^s = 0, \quad b_{ij}^s = 0, \quad \forall i \neq j. \end{aligned} \quad (12)$$

Note that the LS term in the penalized cost only aggregates residuals when  $\hat{\chi}_{\tau s} = 1$  ( $\hat{\sigma}(t) = s$ ). In principle, during interval  $t$ , (P1) updates the estimates of  $\mathbf{A}^s$  and  $\mathbf{B}^s$  only if  $\mathbf{Y}_t$  has been generated by submodel  $s$ . Critical to efficiently tracking the hidden network topologies is the need for a reliable approach to estimate  $\chi_{\tau s}$ . Before developing an efficient tracking algorithm that will be suitable to solve (P1), the rest of this section puts forth criteria for estimating  $\sigma(t)$ .

**Estimation of  $\sigma(t)$ .** Given the most recent estimates  $\{\hat{\mathbf{A}}^s, \hat{\mathbf{B}}^s\}_{s=1}^S$  prior to acquisition of  $\mathbf{Y}_t$ , one can estimate  $\sigma(t)$  by minimizing the a priori error over  $S$  i.e.,

$$\hat{\sigma}(t) = \arg \min_{s \in S} \|\mathbf{Y}_t - \hat{\mathbf{A}}^s \mathbf{Y}_t - \hat{\mathbf{B}}^s \mathbf{X}\|_F \quad (13)$$

followed by solving (P1) with  $s = \hat{\sigma}(t)$ . If network dynamics arise from switching between static or slowly-varying per-state topologies, (13) yields a reliable estimate of the most likely state sequence over time.

Alternatively, one may resort to a criterion that depends on minimizing the a posteriori error. This entails first solving (P1) for all states  $s \in S$  upon acquisition of  $\mathbf{Y}^t$ , and then selecting  $s = \hat{\sigma}(t)$  so that

$$\hat{\sigma}(t) = \arg \min_{s \in S} \|\mathbf{Y}_t - \hat{\mathbf{A}}^{(s|\sigma(t)=s)} \mathbf{Y}_t - \hat{\mathbf{B}}^{(s|\sigma(t)=s)} \mathbf{X}\|_F \quad (14)$$

where  $\hat{\mathbf{A}}^{(s|\sigma(t)=s)}$  (resp.  $\hat{\mathbf{B}}^{(s|\sigma(t)=s)}$ ) denotes the estimate of  $\mathbf{A}^s$  (resp.  $\mathbf{B}^s$ ) given that  $\sigma(t) = s$ . In the context of big data and online streaming, (14) is prohibitive, and (13) is more desirable for its markedly lower computational overhead. The tracking algorithm developed next adopts (13) for state

sequence estimation.

## V. TOPOLOGY TRACKING ALGORITHM

In order to solve (P1), this section resorts to proximal gradient (PG) approaches, which have been popularized for  $\ell_1$ -norm regularized problems, through the class of iterative shrinkage-thresholding algorithms (ISTA); see e.g., [6] and [17]. Unlike off-the-shelf interior point methods, ISTA is computationally simple, with iterations entailing matrix-vector multiplications, followed by a soft-thresholding operation [9, p. 93]. Motivated by its well-documented merits, an ISTA algorithm is developed in the sequel for recursively solving (P1) per  $t$ . Memory storage and computational costs incurred by the algorithm per acquired sample  $\mathbf{Y}_t$  do not grow with  $t$ .

**Solving (P1) for a single time interval  $t$ .** Introducing the optimization variable  $\mathbf{V}^s := [\mathbf{A}^s \quad \mathbf{B}^s]$ , it follows that the gradient of  $f(\mathbf{V}^s) := (1/2) \sum_{\tau=1}^t \hat{\chi}_{\tau s} \|\mathbf{Y}_\tau - \mathbf{A}^s \mathbf{Y}_\tau - \mathbf{B}^s \mathbf{X}\|_F^2$  is Lipschitz continuous, meaning there exists a constant  $L_f$  so that  $\|\nabla f(\mathbf{V}_1^s) - \nabla f(\mathbf{V}_2^s)\| \leq L_f \|\mathbf{V}_1^s - \mathbf{V}_2^s\|, \forall \mathbf{V}_1^s, \mathbf{V}_2^s$  in the domain of  $f$ . Instead of directly optimizing the cost in (P1), PG algorithms minimize a sequence of overestimators evaluated at the current iterate, or a linear combination of the two previous iterates.

Letting  $k = 1, 2, \dots$  denote the iteration index, and  $g(\mathbf{V}^s) := \lambda_s \|\mathbf{A}^s\|_1$ , PG algorithms iterate

$$\mathbf{V}^s[k] := \arg \min_{\mathbf{V}} \left\{ \frac{L_f}{2} \|\mathbf{V} - \mathbf{G}(\mathbf{V}^s[k-1])\|_F^2 + g(\mathbf{V}) \right\} \quad (15)$$

where  $\mathbf{G}(\mathbf{V}^s[k-1]) := \mathbf{V}^s[k-1] - (1/L_f) \nabla f(\mathbf{V}^s[k-1])$  corresponds to a gradient-descent step taken from  $\mathbf{V}^s[k-1]$ , with step-size equal to  $1/L_f$ . The optimization problem (15) is denoted by  $\text{prox}_{g/L_f}(\mathbf{G}(\mathbf{V}^s[k-1]))$ , and is known as the proximal operator of the function  $g/L_f$  evaluated at  $\mathbf{G}(\mathbf{V}^s[k-1])$ . With  $\mathbf{G}^s[k-1] := \mathbf{G}(\mathbf{V}^s[k-1])$  for convenience, the PG iterations can be compactly rewritten as  $\mathbf{V}[k] = \text{prox}_{g/L_f}(\mathbf{G}^s[k-1])$ .

The success of PG algorithms hinges upon efficient evaluation of the proximal operator (cf. (15)). Focusing on (P1), note that (15) decomposes into

$$\begin{aligned} \mathbf{A}^s[k] &:= \arg \min_{\mathbf{A}^s} \left\{ \frac{L_f}{2} \|\mathbf{A}^s - \mathbf{G}_A^s[k-1]\|_F^2 + \lambda_s \|\mathbf{A}^s\|_1 \right\} \\ &= \text{soft}_{\lambda_s/L_f}(\mathbf{G}_A^s[k-1]) \\ \mathbf{B}^s[k] &:= \arg \min_{\mathbf{B}^s} \left\{ \|\mathbf{B}^s - \mathbf{G}_B^s[k-1]\|_F^2 \right\} = \mathbf{G}_B^s[k-1] \end{aligned} \quad (16)$$

(17)

subject to the constraints in (P1) which so far have been left implicit, and  $\mathbf{G}^s := [\mathbf{G}_A^s \quad \mathbf{G}_B^s]$ . Letting  $\text{soft}_\mu(\mathbf{M})$  with  $(i, j)$ -th entry given by  $\text{sign}(m_{ij}) \max(|m_{ij}| - \mu, 0)$  denote the soft-thresholding operator, it follows that  $\text{prox}_{\lambda_s \|\cdot\|_1/L_f}(\cdot) = \text{soft}_{\lambda_s/L_f}(\cdot)$ , see e.g., [6], [9]. Since there is no regularization on  $\mathbf{B}^s$ , (17) boils-down to a simple gradient-descent step.

Specification of  $\mathbf{G}_A^s$  and  $\mathbf{G}_B^s$  only requires expressions for the gradient of  $f(\mathbf{V}^s)$  with respect to  $\mathbf{A}^s$  and  $\mathbf{B}^s$ . Note that

by incorporating the constraints  $a_{ii}^s = 0$  and  $b_{ij}^s = 0, \forall j \neq i, i = 1, \dots, N$ , one can express  $f(\mathbf{V}^s)$  as

$$f(\mathbf{V}^s) := \frac{1}{2} \sum_{\tau=1}^t \sum_{i=1}^N \hat{\chi}_{\tau s} \|\mathbf{y}_{i,\tau}^\top - (\mathbf{a}_{-i}^s)^\top \mathbf{Y}_{-i,\tau} - b_{ii}^s \mathbf{x}_i^\top\|_2^2 \quad (18)$$

where  $\mathbf{y}_{i,\tau}^\top$  and  $\mathbf{x}_i^\top$  denote the  $i$ -th row of  $\mathbf{Y}_\tau$  and  $\mathbf{X}$ , respectively; while  $(\mathbf{a}_{-i}^s)^\top$  denotes the  $1 \times (N-1)$  vector obtained by removing entry  $i$  from the  $i$ -th row of  $\mathbf{A}^s$ , and likewise  $\mathbf{Y}_{-i,\tau}$  is the  $(N-1) \times C$  matrix obtained by removing row  $i$  from  $\mathbf{Y}_\tau$ . It is apparent from (18) that  $f(\mathbf{V}^s)$  is separable across the trimmed row vectors  $(\mathbf{a}_{-i}^s)^\top$ , and the diagonal entries  $b_{ii}^s$ ,  $i = 1, \dots, N$ . The sought gradients are

$$\begin{aligned} \nabla_{\mathbf{a}_{-i}^s} f(\mathbf{V}^s) &= \mathbf{\Omega}_{-i,t}^s \mathbf{a}_{-i}^s + \bar{\mathbf{Y}}_{-i,t}^s \mathbf{x}_i b_{ii}^s - \boldsymbol{\omega}_{-i,t}^s \quad (19) \\ \nabla_{b_{ii}^s} f(\mathbf{V}^s) &= (\mathbf{a}_{-i}^s)^\top \bar{\mathbf{Y}}_{-i,t}^s \mathbf{x}_i + \alpha_t^s b_{ii}^s \|\mathbf{x}_i\|_2^2 - (\bar{\mathbf{y}}_{i,t}^s)^\top \mathbf{x}_i \quad (20) \end{aligned}$$

where  $(\bar{\mathbf{y}}_{i,t}^s)^\top$  denotes the  $i$ -th row of  $\bar{\mathbf{Y}}_t^s := \sum_{\tau=1}^t \hat{\chi}_{\tau s} \mathbf{Y}_\tau$ ,  $\bar{\mathbf{Y}}_{-i,t}^s := \sum_{\tau=1}^t \hat{\chi}_{\tau s} \mathbf{Y}_{-i,\tau}$ , and  $\alpha_t^s := \sum_{\tau=1}^t \hat{\chi}_{\tau s}$ . Similarly,  $\boldsymbol{\omega}_{-i,t}^s := \sum_{\tau=1}^t \hat{\chi}_{\tau s} \mathbf{Y}_{-i,\tau} \mathbf{y}_{i,\tau}$  and  $\mathbf{\Omega}_{-i,t}^s$  is obtained by removing the  $i$ -th row and  $i$ -th column from  $\mathbf{\Omega}_t^s := \sum_{\tau=1}^t \hat{\chi}_{\tau s} \mathbf{Y}_\tau \mathbf{Y}_\tau^\top$ . From (16)-(17) and (19)-(20), the parallel ISTA iterations

$$\begin{aligned} \nabla_{\mathbf{a}_{-i}^s} f[k] &= \mathbf{\Omega}_{-i,t}^s \mathbf{a}_{-i}^s[k] + \bar{\mathbf{Y}}_{-i,t}^s \mathbf{x}_i b_{ii}^s[k] - \boldsymbol{\omega}_{-i,t}^s \quad (21) \\ \nabla_{b_{ii}^s} f[k] &= (\mathbf{a}_{-i}^s)^\top [k] \bar{\mathbf{Y}}_{-i,t}^s \mathbf{x}_i + \alpha_t^s b_{ii}^s[k] \|\mathbf{x}_i\|_2^2 - (\bar{\mathbf{y}}_{i,t}^s)^\top \mathbf{x}_i \quad (22) \end{aligned}$$

$$\mathbf{a}_{-i}^s[k+1] = \text{soft}_{\lambda_s/L_f} \left( \mathbf{a}_{-i}^s[k] - (1/L_f) \nabla_{\mathbf{a}_{-i}^s} f[k] \right) \quad (23)$$

$$b_{ii}^s[k+1] = b_{ii}^s[k] - (1/L_f) \nabla_{b_{ii}^s} f[k] \quad (24)$$

are provably convergent to the globally optimal solution  $\{\hat{\mathbf{A}}^s, \hat{\mathbf{B}}^s\}$  of (P1), as per general convergence results for PG methods [6], [17].

Note that iterations (23)-(24) incur a low computational overhead, involving at most matrix-vector multiplication complexity for the gradient evaluations. A final step entails zero-padding the updated  $\mathbf{a}_{-i}^s[k]$  by setting

$$(\mathbf{a}_i^s)^\top [k] = [a_{-i,1}^s[k] \dots a_{-i,i-1}^s[k] \ 0 \ a_{-i,i+1}^s[k] \dots a_{-i,N}^s[k]]. \quad (25)$$

The desired SEM parameter estimates are subsequently obtained as  $\hat{\mathbf{A}}^s = [(\mathbf{a}_1^s)^\top [k], \dots, (\mathbf{a}_N^s)^\top [k]]^\top$  and  $\hat{\mathbf{B}}^s = \text{diag}(b_{11}^s[k], \dots, b_{NN}^s[k])$ , for  $k$  large enough so that convergence has been attained.

**Solving (P1) over the entire time horizon.** Tracking the switching sequence and the state parameters entails sequentially alternating between two operations per datum arrival. First,  $\hat{\sigma}(t)$  is estimated (via solving (13) or (14)), and the corresponding values  $\{\hat{\chi}_{ts}\}_{s=1}^S$  are accordingly obtained. The iteration steps (21)-(24) are then run until convergence is attained. Note that these iterations only depend on past data through recursively updated moving averages, that is,

$$\mathbf{\Omega}_t^s = \mathbf{\Omega}_{t-1}^s + \hat{\chi}_{ts} \mathbf{Y}_t \mathbf{Y}_t^\top, \quad \bar{\mathbf{Y}}_t^s = \bar{\mathbf{Y}}_{t-1}^s + \hat{\chi}_{ts} \mathbf{Y}_t. \quad (26)$$

Similar recursive expressions can be readily derived for  $\bar{\mathbf{Y}}_{-i,t}^s$ ,  $\alpha_t^s$ , and  $\boldsymbol{\omega}_{-i,t}^s$ . The complexity in evaluating the Gram

---

### Algorithm 2 Topology tracking algorithm

---

**Require:**  $\{\mathbf{Y}_t\}_{t=1}^T, \mathbf{X}, S, \{\lambda_s\}_{s=1}^S$   
1: Initialize  $\{\hat{\mathbf{A}}_0^s, \hat{\mathbf{B}}_0^s, \bar{\mathbf{Y}}_0^s = \mathbf{0}_{N \times C}\}_{s=1}^S$   
2: **for**  $t = 1, \dots, T$  **do**  
3:    $\hat{\sigma}(t) = \arg \min_{s \in \mathcal{S}} \|\mathbf{Y}_t - \hat{\mathbf{A}}_{t-1}^s \mathbf{Y}_t - \hat{\mathbf{B}}_{t-1}^s \mathbf{X}\|_F$   
4:   Set  $\hat{\chi}_{t\hat{\sigma}(t)} = 1$  and  $\hat{\chi}_{ts} = 0$  for  $s \in \mathcal{S} \setminus \hat{\sigma}(t)$   
5:   **for**  $s = 1, \dots, S$  **do**  
6:     **if**  $s = \hat{\sigma}(t)$  **then**  
7:       Update  $\mathbf{\Omega}_t^s, \bar{\mathbf{Y}}_t^s, \alpha_t^s, L_f$   
8:       Set  $\mathbf{A}^s[0] = \hat{\mathbf{A}}_{t-1}^s, \mathbf{B}^s[0] = \hat{\mathbf{B}}_{t-1}^s, k = 0$   
9:       **while** not converged **do**  
10:          **for**  $i = 1 \dots N$  (in parallel) **do**  
11:            $\mathbf{z}^s[k] = \mathbf{a}_{-i}^s[k] - (1/L_f) \nabla_{\mathbf{a}_{-i}^s} f[k]$   
12:            $\mathbf{a}_{-i}^s[k+1] = \text{soft}_{\lambda_s/L_f}(\mathbf{z}^s[k])$   
13:            $b_{ii}^s[k+1] = b_{ii}^s[k] - (1/L_f) \nabla_{b_{ii}^s} f[k]$   
14:           Update  $\mathbf{a}_i^s[k+1]$  via (25)  
15:          **end for**  
16:           $k = k + 1$   
17:       **end while**  
18:        $\hat{\mathbf{A}}_t^s = \mathbf{A}^s[k], \hat{\mathbf{B}}_t^s = \mathbf{B}^s[k]$   
19:     **else**  
20:        $\hat{\mathbf{A}}_t^s = \hat{\mathbf{A}}_{t-1}^s, \hat{\mathbf{B}}_t^s = \hat{\mathbf{B}}_{t-1}^s$   
21:        $\alpha_t^s = \alpha_{t-1}^s, \mathbf{\Omega}_t^s = \mathbf{\Omega}_{t-1}^s, \bar{\mathbf{Y}}_t^s = \bar{\mathbf{Y}}_{t-1}^s$   
22:     **end if**  
23:   **end for**  
24: **end for**  
25: **return**  $\{\hat{\mathbf{A}}_t^s, \hat{\mathbf{B}}_t^s\}_{s=1}^S, \{\hat{\sigma}(t)\}_{t=1}^T$

---

matrix  $\mathbf{Y}_t \mathbf{Y}_t^\top$  dominates the per-iteration computational cost of the algorithm. The recursive updates in (26) are conducted only for a single state  $s = \hat{\sigma}(t)$  per interval  $t$ , with the remainder  $\{\mathbf{\Omega}_t^s, \bar{\mathbf{Y}}_t^s\}_{s \in \mathcal{S} \setminus \hat{\sigma}(t)}$  set to  $\{\mathbf{\Omega}_{t-1}^s, \bar{\mathbf{Y}}_{t-1}^s\}_{s \in \mathcal{S} \setminus \hat{\sigma}(t)}$ . Similarly, iterations (21)-(24) are run only for  $s = \hat{\sigma}(t)$ , while  $\{\hat{\mathbf{A}}^s, \hat{\mathbf{B}}^s\}_{s \in \mathcal{S} \setminus \hat{\sigma}(t)}$  are not updated during interval  $t$ . Furthermore, the need to recompute  $L_f$  per  $t$  can be circumvented by selecting an appropriate step-size for (23)-(24) by line search [17].

Algorithm 2 summarizes the developed state-dependent topology tracking scheme. Numerical tests indicate that 3 – 5 inner ISTA iterations suffice to track the evolving topology remarkably well.

#### A. Initialization of Algorithm 2

In order to run Algorithm 2, one needs initial state estimates  $\{\mathbf{A}_0^s, \mathbf{B}_0^s\}_{s=1}^S$ . If  $\mathbf{X}$  is known to have full row rank, Algorithm 1 can be used to initialize the state matrices as the  $S$  cluster centroids. However, this is quite restrictive, and a more general initialization scheme can be obtained with less stringent restrictions on  $\mathbf{X}$ . For example, the following regularized LS estimator with  $\mu > 0$

$$\arg \min_{\{\mathbf{A}, \mathbf{B}: a_{ii}=0, b_{ij}=0\}} (1/2) \|\mathbf{Y}_t - \mathbf{A} \mathbf{Y}_t - \mathbf{B} \mathbf{X}\|_F^2 + \mu \|\mathbf{A}\|_F^2 \quad (27)$$

yields estimates  $\{\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t\}_{t=1}^{T_{\text{init}}}$  per  $t$  for a designated initialization interval  $t = 1, \dots, T_{\text{init}}$  ( $T_{\text{init}} \ll T$ ). The initializations  $\{\mathbf{A}_0^s, \mathbf{B}_0^s\}_{s=1}^S$  are then obtained as the  $S$  cluster centroids

of  $\{\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t\}_{t=1}^{T_{\text{init}}}$ . Note that (27) decouples across nodes, and amounts to solving

$$\arg \min_{\mathbf{a}_{-i}, b_{ii}} (1/2) \|\mathbf{y}_{i,t} - \mathbf{Y}_{-i,t}^\top \mathbf{a}_{-i} - b_{ii} \mathbf{x}_i\|_2^2 + \mu \|\mathbf{a}_{-i}\|_2^2 \quad (28)$$

for  $i = 1, \dots, N$ . Indeed, (28) admits the following per-variable closed-form solutions

$$\mathbf{a}_{-i} = (\mathbf{Y}_{-i,t} \mathbf{Y}_{-i,t}^\top + 2\mu \mathbf{I})^{-1} \mathbf{Y}_{-i,t} (\mathbf{y}_{i,t} - b_{ii} \mathbf{x}_i) \quad (29)$$

and

$$b_{ii} = \frac{(\mathbf{a}_{-i}^\top \mathbf{Y}_{-i,t} - \mathbf{y}_{i,t})^\top \mathbf{x}_i}{\|\mathbf{x}_i\|_2^2} \quad (30)$$

for  $i = 1, \dots, N$ . Starting with an initial value for  $b_{ii}$ , one can compute (29) and (30) in an alternating fashion by fixing one variable and updating the other, until convergence is attained.

### B. Tracking slowly-changing state matrices

It has tacitly been assumed that state matrices  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  are static, and that temporal dynamics only arise from the switching sequence  $\{\sigma(t)\}_{t=1}^T$ . Nevertheless, entries of each of the state matrices may drift slowly over time, motivating algorithms that down-weight the influence of past data. To this end, (P1) can be modified as follows

$$\begin{aligned} \text{(P2)} \quad \arg \min_{\mathbf{A}^s, \mathbf{B}^s} \quad & \frac{1}{2} \sum_{\tau=1}^t \beta^{t-\tau} \hat{\chi}_{\tau s} \|\mathbf{Y}_\tau - \mathbf{A}^s \mathbf{Y}_\tau - \mathbf{B}^s \mathbf{X}\|_F^2 \\ & + \lambda_s \|\mathbf{A}^s\|_1 \\ \text{s. to} \quad & a_{ii}^s = 0, b_{ij}^s = 0, \forall i \neq j. \end{aligned} \quad (31)$$

where  $\beta \in (0, 1]$  is a forgetting factor that exponentially down-weights past data whenever  $\beta < 1$ . In order to solve (P2), Algorithm 2 can be readily modified in a reasonably straightforward manner, and details are omitted here for brevity.

## VI. NUMERICAL TESTS

### A. Synthetic Data

**Data generation.** To assess the performance of the developed algorithms, the first set of experiments were conducted on synthetic cascade data. Four Kronecker graphs were generated from the following seed matrices [11]

$$\begin{aligned} \mathbf{H}_1 &= \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} & \mathbf{H}_2 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \\ \mathbf{H}_3 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} & \mathbf{H}_4 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

The resulting graph adjacency matrices  $\{\mathbf{A}^s \in \mathbb{R}^{64 \times 64}\}_{s=1}^4$ , each encoding a network of  $N = 64$  nodes, were obtained by repeated Kronecker products  $\mathbf{A}^s = \mathbf{H}_s \otimes \mathbf{H}_s \otimes \mathbf{H}_s$  for  $s = 1, \dots, 4$ . Each diagonal entry of  $\mathbf{A}^s$  was set to zero, while the number of cascades was set to  $C = 80$ , and time intervals to  $T = 1,000$ . Furthermore,  $\mathbf{X} \in \mathbb{R}^{N \times C}$  was constructed with entries sampled from a uniform distribution as  $[\mathbf{X}]_{ij} \sim \mathcal{U}[0, 3]$ . Similarly, the diagonal matrices  $\{\mathbf{B}^s \in \mathbb{R}^{N \times N}\}_{s=1}^4$  were

constructed by sampling entries from a uniform distribution as  $[\mathbf{B}]_{ii} \sim \mathcal{U}[0, 1]$ . Synthetic cascade data were then generated as  $\mathbf{Y}_t = (\mathbf{I}_N - \mathbf{A}^{\sigma(t)})^{-1}(\mathbf{B}^{\sigma(t)} \mathbf{X} + \mathbf{E}_t)$ , with  $\sigma(t)$  sampled uniformly at random from  $S = \{1, 2, 3, 4\}$  per  $t$ , and  $[\mathbf{E}_t]_{ij} \sim \mathcal{N}(0, 0.01)$ , for  $t = 1, \dots, T$ . Figure 3 (a) depicts the ground-truth adjacency matrices  $\{\mathbf{A}^s\}_{s=1}^4$  used to generate the synthetic cascade data.

**Experimental results.** First, Algorithm 1 was run with  $T_{\text{train}} = 200$ , and k-means as the clustering algorithm of choice. Figure 3 (b) depicts heatmaps of the resulting adjacency matrices obtained as cluster centroids. Note that Algorithm 1 is devoid of a data-driven thresholding scheme, hence most entries in the recovered adjacency matrix are nonzero. Nevertheless, visual inspection of the heatmaps reveals that larger entries generally correspond to non-zero edge weights as shown in Figure 3 (a). Since the synthetic cascade data are noisy, it is not surprising that Algorithm 1 exhibits suboptimal topology identification. In fact as demonstrated next, Algorithm 2 is markedly superior to Algorithm 1 with respect to coping with noise, as well as shrinking non-edge entries of  $\{\mathbf{A}^s\}_{s=1}^4$  to zero.

Using the initial 50 cascade samples  $\{\mathbf{Y}_t\}_{t=1}^{50}$ , per-interval state-agnostic batch estimates  $\{\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t\}_{t=1}^{50}$  were obtained by solving (27) with  $\mu = 0.01$ . The estimates  $\{\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t\}_{t=1}^{50}$  were then clustered via the k-means algorithm, with  $k = 4$  clusters, and the corresponding cluster centers were used as the initial values  $\{\hat{\mathbf{A}}_0^s, \hat{\mathbf{B}}_0^s\}_{s=1}^4$ . Algorithm 2 was then run for  $t = 51, \dots, 1,000$ , with  $\lambda_s = 0.95$ , and  $s \in S$ . Figure 3 (c) depicts heatmaps of the recovered state-dependent adjacency matrices at  $T = 1,000$ . Thanks to sparsity regularization, Algorithm 2 unveils the non-zero support structures of the state matrices with remarkable success. Figure 4 plots the actual and estimated switching sequences from  $t = 901$  to  $t = 1,000$ , clearly demonstrating the remarkable success in tracking the underlying states.

Next, the advocated approach was compared with earlier work which models temporal information cascades using dynamic SEMs, with slow topology variations [1]. A stochastic gradient descent (SGD) algorithm developed in [1] was run with batch initialization on the time-series of cascade data. Per interval  $t$ , the resulting relative estimation error,  $(\|\mathbf{A}^t - \hat{\mathbf{A}}^t\|_F + \|\mathbf{B}^t - \hat{\mathbf{B}}^t\|_F) / (\|\hat{\mathbf{A}}^t\|_F + \|\hat{\mathbf{B}}^t\|_F)$ , was evaluated for both Algorithm 2 and the SGD tracker from [1]. Note that in computing the estimation error for Algorithm 2,  $\mathbf{A}^t = \mathbf{A}^{\sigma(t)}$ ,  $\mathbf{B}^t = \mathbf{B}^{\sigma(t)}$  (similarly for  $\hat{\mathbf{A}}^t$  and  $\hat{\mathbf{B}}^t$ ). Figure 5 (a) compares the per-interval relative errors of the two approaches. It is clear from the plot that exploiting the prior knowledge that network dynamics arise due to random switching between  $S = 4$  states yields remarkably superior error performance than the alternative. In fact, the final adjacency matrix  $\hat{\mathbf{A}}^T$  does not correspond to a specific state, but is rather an average of the underlying state matrices. However, this is not surprising since the framework advocated in [1] exploits slow topology variations, and it is not expected to outperform algorithms developed in the present paper, when topologies potentially jump suddenly between discrete states.

To this end, a new slowly-varying state sequence was used to control the evolving network topologies, and a new time series of synthetic cascades was generated. Specifically, the



following scheme was used to generate a piecewise-constant sequence  $\{\sigma(t)\}_{t=1}^T$ :

$$\sigma(t) = \begin{cases} 1, & t \in \{1, \dots, 24\} \cup \{200, \dots, 299\} \\ 2, & t \in \{25, \dots, 49\} \cup \{300, \dots, 699\} \\ 3, & t \in \{50, \dots, 74\} \cup \{700, \dots, 899\} \\ 4, & t \in \{75, \dots, 199\} \cup \{900, \dots, 1,000\} \end{cases}$$

while the ground-truth matrices  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  and  $\{\mathbf{E}_t\}_{t=1}^T$  remained unchanged. Figure 5 (a) compares per-interval relative errors resulting from running Algorithm 2, and the SGD algorithm developed in [1]. In this case, there are fewer sudden transitions between network states. Nevertheless, leveraging the prior information about the state-based network evolution is beneficial as shown by the plot. Indeed, modeling cascade propagation by the proposed switched dynamic SEM framework leads to better error performance than adopting a state-incognizant approach that exploits the underlying slow variations.

### B. Real Cascade Data

**Dataset description.** This section presents results of experimental tests conducted on real cascades observed on the web between March 2011 and February 2012. Blog posts and news articles for memes (popular textual phrases) appearing within on approximately 3.3 million websites were monitored during the observation period. The time when a given website first mentioned a story related to a specific set of memes was recorded, and the data were availed to the public from [InfoPath]. Cascade infection times were recorded as Unix timestamps in hours (i.e., the number of hours since midnight on January 1, 1970). Several trending news topics during this period were identified, and cascade data for the top 5,000 websites that mentioned memes associated with them were retained. From this dataset, cascades related to 10 broad news topics were extracted for the present paper's experiments. Table I lists the names of these topics, with a correspondingly brief description.

For each broad topic in Table I, many news stories appeared on blogs and mainstream news websites over the observation period. Each story was assigned a list of tuples in the form (website id, timestamp) capturing the time when it was first mentioned on a particular website. For the present paper, a popular news story was characterized as a meme if it propagated to at least 100 websites, due to repeated mentions. Based on this definition, all news stories that did not qualify as memes were discarded, and a total of  $C = 625$  cascades constituting the selected memes, and their infection times were retained. Due to this thresholding, the total number of websites reduced to  $N = 1,131$ . The observation period was split into  $T = 180$  time intervals  $\{\mathcal{T}_t\}_{t=1}^{180}$ , each corresponding to approximately 2 days. Letting  $u_{ic}$  denote the Unix infection time of node  $i$  by cascade  $c$ , datum  $y_{ic}^t$  was computed as follows:

$$y_{ic}^t = \log_{10} \left( u_{ic} - \min_{u_{ic} \in \mathcal{T}_t} u_{ic} \right) \quad (32)$$

only if  $u_{ic} \in \mathcal{T}_t$ , otherwise  $y_{ic}^t$  is set to a very large value, namely  $y_{ic}^t = 2 + \log_{10}(\max_{i,c} u_{ic})$ , as a surrogate for infinity.

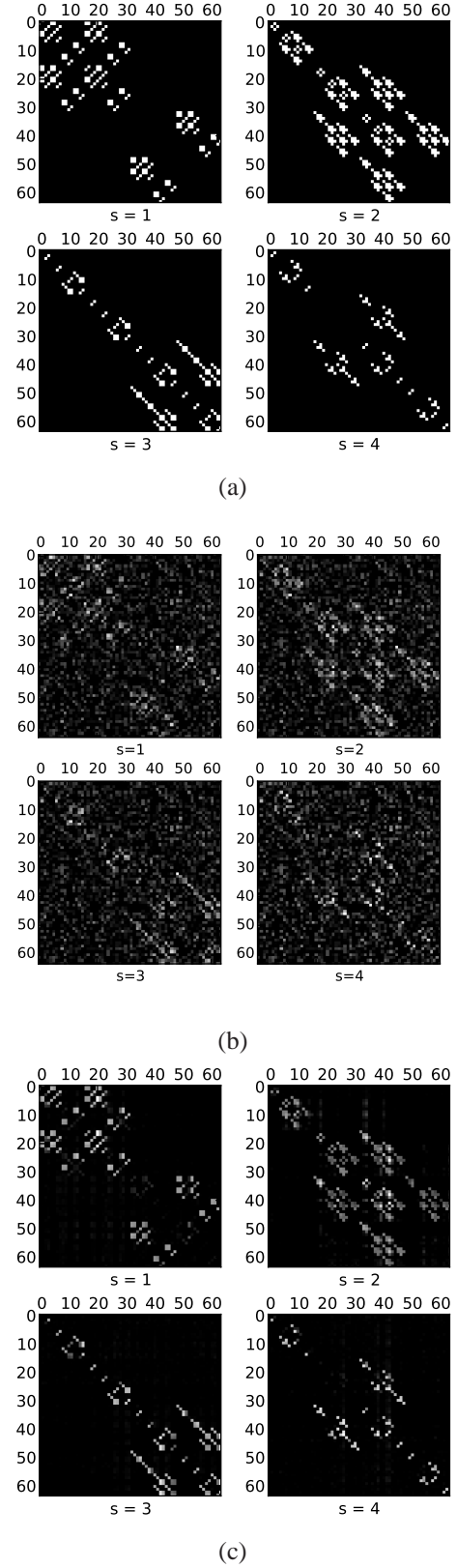


Fig. 3: Adjacency matrices corresponding to: (a) actual switching networks used for the synthetic dataset; (b) state-dependent network topologies inferred by Algorithm 1; and (c) state-dependent network topologies inferred by Algorithm 2.



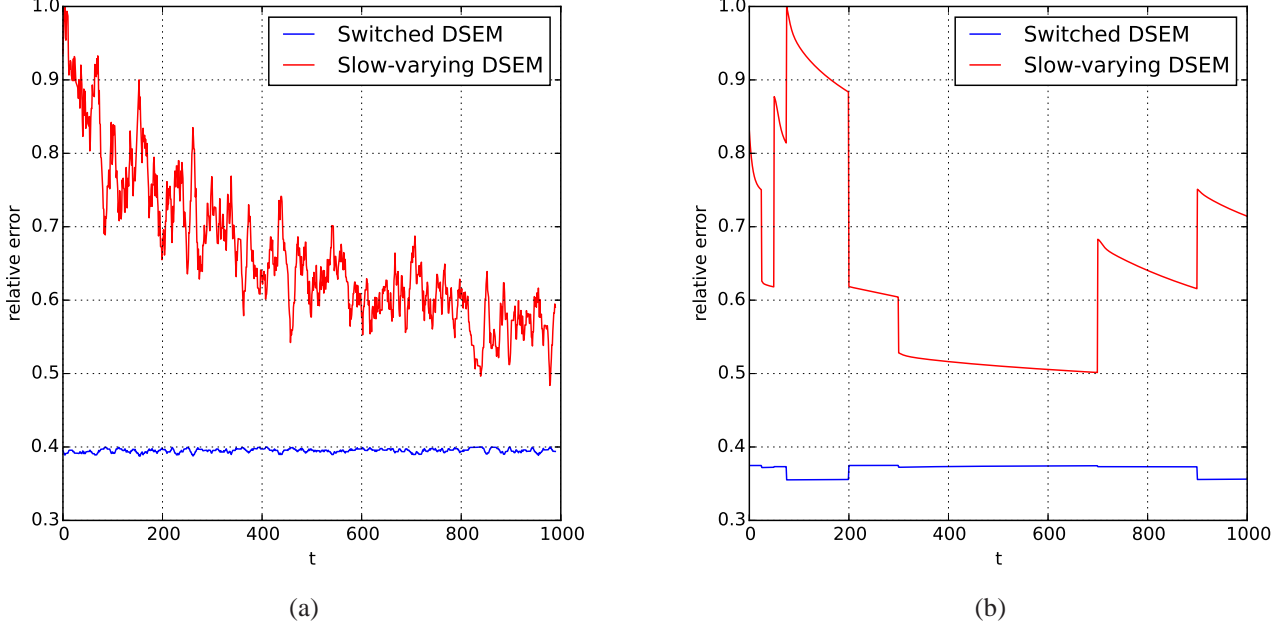


Fig. 5: Comparison of relative errors resulting from tracking the unknown topologies as a switched sequence versus slowly-varying changes: (a) rapidly switching state sequence; and (b) piecewise-constant state sequence.

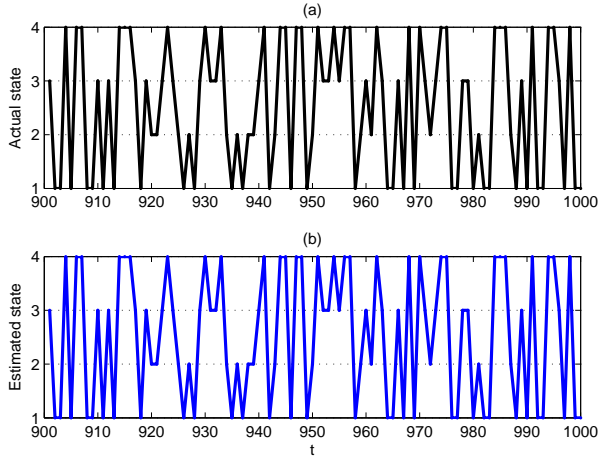


Fig. 4: Actual (a) and estimated (b) switching sequences plotted from  $t = 900$  to  $t = 1,000$ .

Entries of  $\mathbf{X}$  capture prior knowledge about the susceptibility of each node to each contagion. For instance, the entry  $x_{ic}$  could denote the online search rank of website  $i$  for a search keyword associated with contagion  $c$ . In the absence of such website ranking information, the present paper resorted to an alternative approach involving assignment of susceptibilities from the entire corpus of the cascade data. First, five broad categories of memes were identified as *politics*, *entertainment*, *sports*, *business*, and *technology*, and indexed by  $k = 1, \dots, 5$ . Each group of news topics was manually labeled using these five prescribed categories. Next, for each website  $i$ ,  $\gamma_{ik}$  was

computed as follows

$$\gamma_{ik} := \frac{\text{number of category } k \text{ cascades that infected node } i}{\text{total number of cascades that infected node } i}$$

for all  $k$  categories. Entry  $x_{ic}$  was then set to  $\gamma_{ik}$  if cascade  $c$  belonged to category  $k$ .

**Experimental results.** Since  $S$  is unknown for the real dataset, the first part of this experimental test entailed approximating its value from the data. Setting  $\mu = 0.15$ , the initialization scheme in (27) was run in batch mode over 60 intervals. The resulting estimates of  $\{\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t\}_{t=1}^{60}$  were then clustered using  $S$ -means, with  $S = 1, \dots, 10$ . For each  $S$ , the total intra-cluster distance was evaluated upon convergence as

$$\delta(S) := \log_{10} \left\{ \sum_{s=1}^S \sum_{t=1}^{60} \hat{\chi}_{ts} \|\hat{\boldsymbol{\theta}}^t - \hat{\boldsymbol{\theta}}^s\|_2^2 \right\} \quad (33)$$

on a logarithmic scale, with the estimates  $\hat{\chi}_{ts}$ ,  $\hat{\boldsymbol{\theta}}^t$ , and  $\hat{\boldsymbol{\theta}}^s$  defined earlier. Figure 7 plots  $\delta(S)$  for  $S = 1, \dots, 10$ , depicting a substantial decrease from  $S = 2$  to  $S = 3$ . Subsequent values of  $S$  do not markedly reduce  $\delta(S)$ . From this plot,  $S = 3$  was selected as the underlying number of network states.

Next, Algorithm 2 was run with the pre-processed cascade data as inputs. Note that Algorithm 2 requires the sparsity-promoting regularization parameters  $\{\lambda_s\}_{s=1}^S$  as inputs. In this experiment, a uniform value was adopted for all states, i.e.,  $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$ . In the absence of ground-truth state topologies, selection of  $\lambda$  is rather challenging, and it remains an open question for future work. Nevertheless, a heuristic approach inspired by typical properties of real-world large-scale networks was adopted for the present paper. Over the last 10–15 years, several studies in network science have unveiled remarkable universal properties that seem to underlie most

	Broad news topic	Brief description
1	Fukushima	Nuclear accident at Japan's Fukushima nuclear power plant in March 2011
2	Kate Middleton	English royal whose wedding took place in April 2011
3	Kim Jong-un	Leader of North Korea who rose to prominence upon the death of his father in December 2011
4	Osama bin Laden	Infamous terrorist leader who was killed in May 2011
5	Amy Winehouse	Famous English singer who died of drug overdose in July 2011
6	Rupert Murdoch	Businessman whose media company was involved in a phone hacking scandal revealed in May 2011
7	Steve Jobs	Technology entrepreneur whose death in October 2011 was followed by many news headlines
8	Arab spring	Wide-spread politically-charged protests among several Arab nations starting in December 2010
9	Strauss Kahn	International Monetary Fund (IMF) director who resigned in May 2011 due to sex assault allegations
10	Reid Hoffman	Founder of LinkedIn whose stock started trading in May 2011

TABLE I: The 10 broad news topics whose memes constituted the cascades tracked for topology inference.

Network state	1	2	3
Average clustering coefficient	0.367	0.216	0.218
Network diameter	18	13	8
Average number of neighbors	134.164	81.132	25.738
Average shortest path length	9.34	7.66	4.51

TABLE II: Simple summary statistics extracted from the inferred network state topologies with  $\lambda = 10$ .

real networks. For example, networks generally exhibit the so-called “small-world” property, and their degree distributions often follow power laws [24].

Acknowledging these as universal laws inherent to most networks, it is possible to constrain the search space of the most likely network topologies uncovered by Algorithm 2. In this experiment,  $\lambda$  was selected in such a way that the average shortest path length between any pair of nodes is  $\mathcal{O}(\log N)$ , which is consistent with the “small-world” phenomenon. With  $N = 1,131$ , the goal was to select  $\lambda$  so that the resulting average shortest path length was within the interval  $[6, 10]$ . For each  $\lambda$ , Algorithm 2 was run and the average shortest path lengths were computed per network state, and then averaged over all states. Figure 8 plots the average path length as  $\lambda$  is varied between 0 and 100. The plot also depicts the corresponding average network diameters, which can also be used as a validation metric. It turns out that setting  $\lambda = 10$  for  $s \in \{1, 2, 3\}$  led to an average shortest path length of approximately 7, which is indicative of an underlying “small-world” property. Table II summarizes the per-state average clustering coefficient, network diameter, average number of neighbors, and the average shortest path length when  $\lambda = 10$ .

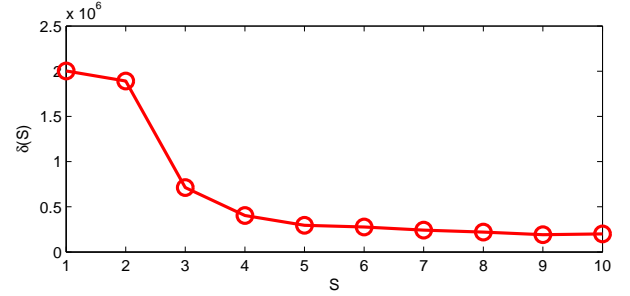
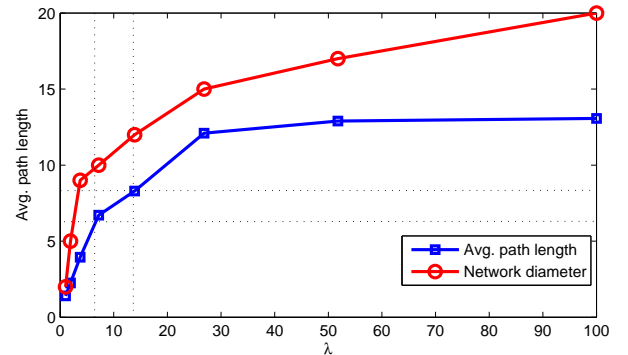
Fig. 7: Total intra-cluster distances plotted for different values of  $S$ , with  $S = 3$  selected as the optimal size of the state space in the real web cascades dataset.Fig. 8: Empirical selection of  $\lambda$  guided by the resulting average shortest path lengths in the inferred networks. With  $\lambda = 10$ , the inferred network topologies exhibit the “small-world” property with average path lengths between 6 – 10 hops.

Figure 6 depicts visualizations of the state-dependent network topologies inferred from the cascade data using Algo-

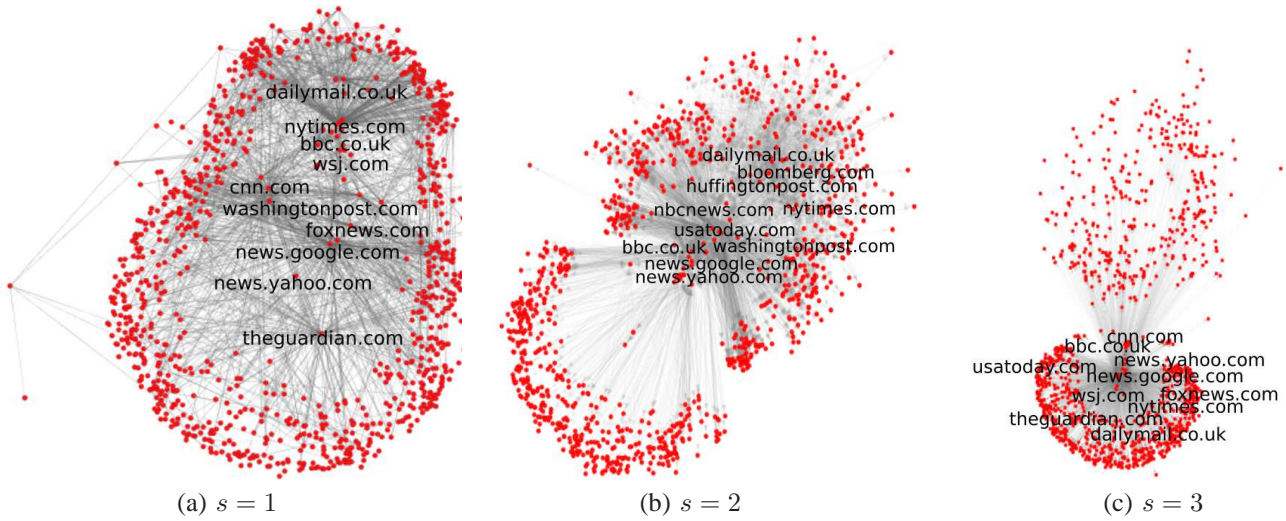


Fig. 6: State-dependent network topologies inferred from real cascades annotated by the top 10 ranked websites, in order of decreasing out-degree. Interestingly, Algorithm 2 reveals that the most influential websites turn out to be well-known media outlets and popular web-based news aggregators.

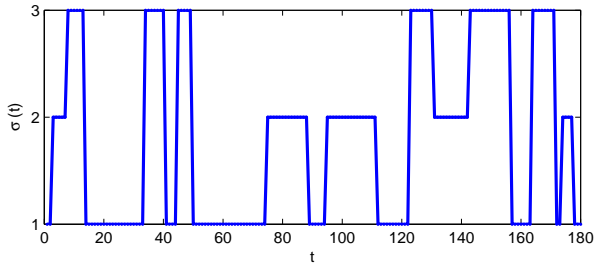


Fig. 9: Inferred switching sequence from the real-world web cascade data. It turns out that network dynamics are governed by piecewise-constant state variations.

rithm 2. Each drawing is annotated with the top 10 websites when ranked in order of decreasing out-degree. It is interesting to note that this list is dominated by websites corresponding to the most influential news outlets in the English-speaking world e.g., *nytimes.com*, *cnn.com*, *bbc.co.uk*, and *theguardian.com*. Furthermore, it turns out that popular news aggregators, such as *news.yahoo.com* and *news.google.com*, are competitive with traditional media powerhouses in terms of influencing the propagation of news over the web. Although this analysis only considers English-based websites, these experimental results on real-world cascades are strongly corroborated by prior expectations about the most influential enablers of information propagation over the world-wide web. Finally, Figure 9 depicts the switching sequence for the network topologies over the entire observation period. It turns out that evolution of the network topology adheres to piecewise constant state variations.

## VII. CONCLUSION

A switched dynamic SEM was proposed for tracking the evolution of dynamic networks from information cascades when topologies arbitrarily switch between discrete states. It was shown that presence of exogenous influences is critical

to guarantee model identifiability, and one can even identify the underlying network topologies with fewer cascade measurements by leveraging edge sparsity. Recognizing that identification of the unknown network states and the switching sequence is computationally challenging, the present paper assumed that the number of states are known a priori, and advocated an exponentially-weighted LS estimator capitalizing on edge sparsity. A recursive two-step topology tracking algorithm leveraging advances in proximal gradient optimization was developed. Per interval, the first step estimates the active state by minimizing the a priori prediction error, while the second step recursively updates only the matrices corresponding to the estimated state.

Experiments on synthetically-generated cascades demonstrated the effectiveness of the advocated approach in jointly tracking the switching sequence, while identifying the state-dependent network topologies. Numerical tests were also conducted on real cascades of popular news over the web, collected over one year. The goal was to identify the underlying topological structure of causal influences between news websites and blogs. Interestingly, websites exhibiting the highest influence (as measured by the out-degree) turned out to be recognizable news outlets and popular news aggregators. Future directions include investigation of more efficient approaches for tracking the switching sequence, identification of the number of states, and selection of sparsity-promoting parameters, especially when ground-truth network topologies are unknown.

## APPENDIX A PROOF OF PROPOSITION 2

In order to construct the proof for Proposition 2, it is necessary to state and prove Lemmas 1 and 2.

**Lemma 1:** Under (as5), matrix  $(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)$  is invertible with probability one.

**Proof:** Note that  $\det(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)$  is a polynomial whose variables are the nonzero entries of  $\mathbf{A}^{\sigma(t)}$ . Consequently, for

some constants  $\{\alpha_{ij}\}_{i,j}$ , the probability that  $(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)$  is not invertible can be equivalently written as

$$\begin{aligned} \Pr\{\det(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top) = 0\} &= \Pr\{\prod_{i,j}(a_{ij}^{\sigma(t)} - \alpha_{ij}) = 0\} \\ &= \Pr\{\bigcup_{i,j}(a_{ij}^{\sigma(t)} = \alpha_{ij})\} \\ &\leq \sum_{i,j} \Pr(a_{ij}^{\sigma(t)} = \alpha_{ij}). \end{aligned} \quad (34)$$

Since  $a_{ij}^{\sigma(t)}$  is drawn from a continuous distribution (as5), it follows that  $\Pr(a_{ij}^{\sigma(t)} = \alpha_{ij}) = 0$ , implying that  $\Pr\{\det(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top) \neq 0\} = 1$ . ■

**Lemma 2:** Under (as4)-(as6), assuming  $\text{kr}(\mathbf{X}^\top) \geq 2K + 1$ , implies that

$$\text{kr}(\mathbf{Y}_t^\top) = \text{kr}\left(\mathbf{X}^\top \mathbf{B}^{\sigma(t)} \left(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top\right)^{-1}\right) \geq 2K + 1$$

with probability one.

**Proof:** Defining  $\bar{\mathbf{X}}^\top := \mathbf{X}^\top \mathbf{B}^{\sigma(t)}$ , note that  $\text{kr}(\bar{\mathbf{X}}^\top) = \text{kr}(\mathbf{X}^\top \mathbf{B}^{\sigma(t)})$ . Let  $\bar{\mathbf{X}}_Q^\top$  denote a submatrix of  $\bar{\mathbf{X}}^\top$ , formed by selecting an arbitrary subset of columns indexed by the set  $Q \subset \{1, \dots, N\}$  whose cardinality is  $2K + 1$ . Given that  $\text{kr}(\bar{\mathbf{X}}^\top) \geq 2K + 1$ , there exists a set of rows of  $\bar{\mathbf{X}}_Q^\top$  indexed by  $\mathcal{P} \subset \{1, \dots, C\}$  with cardinality  $2K + 1$ , such that the resulting square submatrix  $\bar{\mathbf{X}}_{Q,\mathcal{P}}^\top$  has full rank.

Since from Lemma 1, we have that  $(\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)$  is invertible, it holds that  $\mathbf{Y}_t^\top = \bar{\mathbf{X}}^\top (\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)^{-1}$ , or  $\mathbf{Y}_t^\top = \bar{\mathbf{X}}^\top \Phi$ , where  $\Phi := (\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)^{-1}$ . Collecting the columns and rows of  $\mathbf{Y}_t^\top$  indexed by  $Q$  and  $\mathcal{P}$  into  $(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top$ , it follows that  $(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top = \bar{\mathbf{X}}_{\mathcal{P}}^\top \Phi_Q$ , where  $\bar{\mathbf{X}}_{\mathcal{P}}^\top$  collects rows of  $\bar{\mathbf{X}}^\top$  indexed by  $\mathcal{P}$ , and  $\Phi_Q$  collects columns of  $\Phi$  indexed by  $Q$ . Equivalently, one can write  $(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top = \bar{\mathbf{X}}_{Q,\mathcal{P}}^\top \Phi_{Q,Q}$ , where only the corresponding columns of  $\bar{\mathbf{X}}_{\mathcal{P}}^\top$  and rows of  $\Phi_Q$  are retained.

By Cramer's rule for matrix inversion, entry  $\phi_{ij}$  of  $\Phi$  can be written as  $\psi_{ij}(\mathbf{A}^{\sigma(t)})/\det(\mathbf{I} - \mathbf{A}^{\sigma(t)})$ , where  $\psi_{ij}(\mathbf{A}^{\sigma(t)})$  denotes a polynomial function of degree at most  $(N - 1)$ , with nonzero entries of  $\mathbf{A}^{\sigma(t)}$  as variables. Furthermore, the entries of  $(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top$  are linear combinations of  $N$  entries of  $\Phi_Q$ , while the determinant of  $(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top$  is a polynomial  $\bar{\psi}_{\det}((\mathbf{Y}_{Q,\mathcal{P}}^t)^\top)$  of degree  $2K + 1$ , with its entries as variables. As a result, one can write the determinant of  $(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top$  as

$$\det((\mathbf{Y}_{Q,\mathcal{P}}^t)^\top) = \frac{\bar{\psi}_{\det}(\mathbf{A}^{\sigma(t)})}{\det(\mathbf{I} - \mathbf{A}^{\sigma(t)})^{(2K+1)N}} \quad (35)$$

where  $\bar{\psi}_{\det}(\mathbf{A}^{\sigma(t)})$  is a polynomial of degree at most  $(2K + 1)N(N - 1)$  with nonzero entries of  $\mathbf{A}^{\sigma(t)}$  as variables, and is formed by the composition of the polynomials  $\psi_{ij}(\mathbf{A}^{\sigma(t)})$  and linear combinations of entries of  $\Phi_Q$ .

It can be shown that  $\bar{\psi}_{\det}(\mathbf{A}^{\sigma(t)})$  is not identically zero as follows. Since  $\mathbf{A}^{\sigma(t)} = \mathbf{0}$  implies that  $\Phi = \mathbf{I}$ , then  $\det((\mathbf{Y}_{Q,\mathcal{P}}^t)^\top) = \det(\bar{\mathbf{X}}_{Q,\mathcal{P}}^\top) \neq 0$ , because  $\bar{\mathbf{X}}_{Q,\mathcal{P}}^\top$  is full rank by construction. Consequently,  $\bar{\psi}_{\det}(\mathbf{A}^{\sigma(t)})$  depends on nonzero entries of  $\mathbf{A}^{\sigma(t)}$ , which are drawn from a continuous distribution. Similar to the argument made in the proof of Lemma 1, the probability that  $\bar{\psi}_{\det}(\mathbf{A}^{\sigma(t)}) = 0$  is zero. Since

$(\mathbf{Y}_{Q,\mathcal{P}}^t)^\top$  has full rank, then  $\text{kr}(\mathbf{Y}_t^\top) \geq 2K + 1$  with probability one. ■

**Proposition 2:** Suppose data matrices  $\mathbf{Y}_t$  and  $\mathbf{X}$  adhere to (4) with  $a_{ii}^{\sigma(t)} = 0$ ,  $b_{ii}^{\sigma(t)} \neq 0$ ,  $\forall i, t$ , and  $b_{ij}^{\sigma(t)} = 0$ ,  $\forall i \neq j, t$ . If (as1)-(as6) hold, and  $p_s > 0$ ,  $\forall s$ , then  $\{\mathbf{A}^s, \mathbf{B}^s\}_{s=1}^S$  can be uniquely identified with probability one as  $T \rightarrow \infty$ .

**Proof:** First rewrite (4) as

$$\mathbf{Y}_t^\top \mathbf{F}_t = \mathbf{X}^\top \quad (36)$$

where  $\mathbf{F}_t := (\mathbf{I} - (\mathbf{A}^{\sigma(t)})^\top)(\mathbf{B}^{\sigma(t)})^{-1}$ . Note that both  $\mathbf{A}^{\sigma(t)}$  and  $\mathbf{B}^{\sigma(t)}$  can be uniquely determined from  $\mathbf{F}_t$ . Specifically, with  $a_{ii}^{\sigma(t)} = 0$ ,  $\forall i$ , then  $\text{Diag}(\mathbf{F}_t) = (\mathbf{B}^{\sigma(t)})^{-1}$ , implying that  $\mathbf{B}^{\sigma(t)} = (\text{Diag}(\mathbf{F}_t))^{-1}$  and  $\mathbf{A}^{\sigma(t)} = \mathbf{I} - \mathbf{B}^{\sigma(t)} \mathbf{F}_t^\top$ . As a result, identifiability of  $\mathbf{F}_t$  implies identifiability of both  $\mathbf{A}^{\sigma(t)}$  and  $\mathbf{B}^{\sigma(t)}$ , and the argument is by contradiction.

If there exists  $\bar{\mathbf{F}}_t$  satisfying (36), then  $\mathbf{Y}_t^\top \bar{\mathbf{F}}_t = \mathbf{0}$ , where  $\tilde{\mathbf{F}}_t := \mathbf{F}_t - \bar{\mathbf{F}}_t$ . Since the columns of  $\tilde{\mathbf{F}}_t$  have at most  $2K + 1$  nonzero entries each, and  $\text{kr}(\mathbf{Y}_t^\top) \geq 2K + 1$  (Lemma 2), it follows that  $\tilde{\mathbf{F}}_t = \mathbf{0}$ . Hence,  $\mathbf{F}_t = \bar{\mathbf{F}}_t$  which is a contradiction. Consequently,  $\mathbf{F}_t$  is uniquely identifiable, and so are  $\mathbf{A}^{\sigma(t)}$  and  $\mathbf{B}^{\sigma(t)}$  for all  $t$ .

Recalling that  $\mathbf{A}^{\sigma(t)} = \sum_{s=1}^S \chi_{ts} \mathbf{A}^s$  and  $\mathbf{B}^{\sigma(t)} = \sum_{s=1}^S \chi_{ts} \mathbf{B}^s$ , the unique solution from (36) coincides with a specific pair in the set  $\{(\mathbf{A}^s, \mathbf{B}^s)\}_{s=1}^S$  per interval  $t$ . Following a similar argument to the proof of Proposition 1, recall that  $\chi_{ts}$  are Bernoulli random variables with  $p_s = \Pr(\chi_{ts} = 1)$  per  $t$ . Hence, the number of times state  $s$  is activated over  $T$  time intervals follows a binomial distribution. With  $T_s := \sum_{t=1}^T \chi_{ts}$  denoting the total number of activations of state  $s$  over  $t = 1, \dots, T$ , it turns out that  $\Pr(T_s \geq 1) = 1 - \Pr(T_s = 0) = 1 - (1 - p_s)^T$ . Given that  $\lim_{T \rightarrow \infty} \Pr(T_s \geq 1) = 1$  only if  $p_s > 0$ , all states can be uniquely identified with probability 1 as  $T \rightarrow \infty$ , whenever  $p_s > 0$ ,  $\forall s$ . ■

## REFERENCES

- [1] B. Baingana, G. Mateos, and G. B. Giannakis, "Proximal-gradient algorithms for tracking cascades over social networks," *IEEE J. Sel. Topics Sig. Proc.*, vol. 8, no. 4, pp. 563–575, Aug. 2014.
- [2] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche, "A recursive identification algorithm for switched linear/affine models," *Nonlinear Analysis: Hybrid Systems*, vol. 5, no. 2, pp. 242–253, 2011.
- [3] S. Basu, A. Shojai, and G. Michailidis, "Network Granger causality with inherent grouping structure," *J. of Machine Learning Research*, vol. 16, no. 2, pp. 417–453, Mar. 2015.
- [4] J. A. Bazerque, B. Baingana, and G. B. Giannakis, "Identifiability of sparse structural equation models for directed and cyclic networks," in *Proc. of the 1st IEEE Intl. Conf. on Signal and Information Processing*, Austin, TX, Dec. 2013.
- [5] X. Cai, J. A. Bazerque, and G. B. Giannakis, "Gene network inference via sparse structural equation modeling with genetic perturbations," *PLoS Comp. Biology*, vol. 9, May 2013.
- [6] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, Aug. 2004.
- [7] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
- [8] A. S. Goldberger, "Structural equation methods in the social sciences," *Econometrica*, vol. 40, pp. 979–1001, Nov. 1972.
- [9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.
- [10] D. Kaplan, *Structural Equation Modeling: Foundations and Extensions*. Sage Publications, 2009.



- [11] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Machine Learning Research*, vol. 11, pp. 985–1042, Mar. 2010.
- [12] B. A. Logsdon and J. Mezey, "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations," *PLoS Comp. Biology*, vol. 6, Dec. 2010.
- [13] A. McIntosh and F. Gonzalez-Lima, "Structural equation modeling and its application to network analysis in functional brain imaging," *Human Brain Mapping*, vol. 2, no. 1, pp. 2–22, 1994.
- [14] S. Meyers and J. Leskovec, "On the convexity of latent social network inference," in *Proc. of Neural Information Processing Systems Conf.*, Vancouver, Canada, Feb. 2013.
- [15] B. Muthén, "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators," *Psychometrika*, vol. 49, pp. 115–132, Mar. 1984.
- [16] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems: A tutorial," *European J. of Control*, vol. 13, no. 2, pp. 242–260, 2007.
- [17] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optimization*, vol. 1, pp. 123–231, 2013.
- [18] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.
- [19] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proc. of 28th Intl. Conf. on Machine Learning*, Bellevue, WA, Jul. 2011.
- [20] M. G. Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *Proc. of 6th Intl. Conf. on Web Search and Data Mining*, Rome, Italy, Dec. 2010.
- [21] E. M. Rogers, *Diffusion of Innovations*. Free Press, 1995.
- [22] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming," *Automatica*, vol. 40, no. 1, pp. 37–50, 2004.
- [23] R. Vidal, "Recursive identification of switched arx systems," *Automatica*, vol. 44, no. 9, pp. 2274–2287, 2008.
- [24] D. J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton University Press, 1999.
- [25] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes," in *Proc. of 16th Intl. Conf. on Artificial Intelligence and Statistics*, Scottsdale, AZ, May 2013.